Disease Cluster Investigation and GIS: A New Paradigm?

Geoffrey M Jacquez, MS, PhD*
BioMedware, Ann Arbor, MI

Abstract

Advances in geographic information system (GIS) and database technologies are introducing a new era of disease control and surveillance. GIS has proven "value added" for targeting public health interventions, identifying study cohorts, mapping disease patterns, and assessing exposures. Nonetheless, it is not entirely clear whether GIS can advance epidemiological science by increasing our understanding of disease etiology. As an enabling technology, the microscope was key in elucidating relationships between pathogens and disease, and made possible fundamental public health advances such as the eradication of smallpox. Does GIS hold equal promise? Can GIS mislead as well as inform us? Can we formulate and test epidemiological hypotheses using GIS? And if we can, what role do disease clustering and other pattern-recognition techniques play? This presentation attempts to place GIS and disease clustering techniques within the context of a systematic approach for formulating and testing epidemiological hypotheses. The elucidation of relationships between disease processes and patterns is identified as an important direction for future research.

Keywords: disease clustering, health surveillance

Introduction

Being one of the last speakers affords me a chance to reflect on the talks and discussions of the last few days. What impresses me the most is how much progress has been made. Three or four years ago many of us were grappling with our first geographic information system (GIS) applications; simply creating a map of exposures and health events justified a presentation. Here I've attended talks that far exceed these tentative first steps. Topics representative of how far and fast we have come include spatial Monte Carlo randomization methods for assessing the significance of spatial patterns, Web-based GIS for dealing with data concurrency and data sharing, and integrated GIS systems for health surveillance and decision-making, to name a few. Indeed, we have come far, but there are flies in the ointment.

Perhaps our biggest weakness is that GIS technology leads the science, and at such a basic level that it determines the very questions we ask of our data. As public health professionals we all know that time is a critical component in all epidemiological processes. Exposure must precede disease outcomes; transmission events require contact in time as well as in space; every disease has a latency period; and so on. Yet time was given little attention in the presentations I've seen at this conference. Why? Because GIS technology leads the science, and time-GIS is not yet commercially available. I think our inability to conduct true space-time queries is one of the greatest

*Geoffrey Jacquez, BioMedware, 516 North State St., Ann Arbor, MI 48104 USA; (p) 734-913-1098; (f) 734-913-2201; E-mail: Jacquez@Biomedware.com

technological deficiencies limiting GIS in public health. It won't be solved until true time-GIS are available. There are other examples of GIS technology leading the science, but I won't go in to them now. What we need is for public health as a science to lead the technology. This will require thinking "outside the box" on our part to identify those epidemiologically valuable functions that are absent from GIS, and incorporation of our input in software development. I believe this is one of the key problems limiting advances in GIS in public health.

Standing here I feel like the Pope preaching to the choir. This conference is attended only by the converted—if you believed GIS were humbug would you be here? Probably not. Is there anyone here who thinks GIS is humbug?[1] Being surrounded by like thinkers can be dangerous. Allow me to play the devil's advocate as I offer some point and counterpoint on statements and observations made over the last few days.

A Dialog with the Devil's Advocate

One of the observations made at this conference is that "all data are spatial," and I think most of us would agree. However, our devil's advocate is a classical epidemiologist, with little or no training in spatial thought. Her counterpoint is, "So what—location is a lousy exposure surrogate." This counterpoint is difficult to parry when we acknowledge that exposure is best measured at the level of the individual.

In the opening plenary session, one of the speakers observed that "the power of GIS is limited only by your imagination," and most of us nodded in agreement. The counterpoint from the devil's advocate is that "it is the expense of GIS, and not its power, that is beyond imagination." And in fact, we all know that establishing a GIS and its data is resource-intensive.

My point is that as GIS enthusiasts we tend not to hear the counterpoint from the devil's advocate. To illustrate: consider two quotes describing different visions of GIS. The first, from David Gerlernter, sees GIS as a powerful representation of our spatial world, depicting the complexity of the ever-changing space in which we live:

> Someday soon you will look into a computer screen and see reality. Some part of your world—the town you live in, the company you work for, your school system, the city hospital—will hang there in a sharp color image, abstract but recognizable, moving subtly in a thousand places. This Mirror World you are looking at is fed by a steady rush of new data pouring in through cables. It is infiltrated by your own software creatures, doing your own business. (1)

This vision is the logical extension of advances now being made in GIS, including self-organizing maps, Web-based GIS, real-time acquisition of Global Positioning System (GPS) data, and open standards allowing access to diverse databases with ready incorporation of "software creatures." In short, Gerlernter envisions GIS as a powerful, enabling technology whose potential in public health is vast and far-reaching. This is consistent with the vision Jack Dangermond presented at lunch yesterday. Contrast this with a second, briefer quote from Marbury, who focuses specifically on the value of GIS in health:

[1] When asked this question, only two in the audience raised their hands.

> For the most part, advances in environmental epidemiology will require carefully designed studies of rigorously defined outcomes combined with good measurements of personal exposure. It would be a shame to be distracted from this effort by the availability of a new tool that affords no new insights. (2)

Marbury recognizes the conundrum facing public health workers: when deciding to undertake one activity, we necessarily commit resources that might have been better spent elsewhere. Such opportunity costs can be substantial for health GIS. Questions such as "Is it wise to spend our health dollars on GIS when we could be vaccinating children?" are powerful illustrations, but of course apply to all public health activities, not just GIS. Here, Marbury is concerned with the opportunity cost of GIS as an epidemiological tool.

Ken Rothman (3) posed similar concerns regarding disease cluster investigations. He observed that cluster investigations usually lead to negative results, are prone to pre-selection bias (the well-known "Texas sharpshooter problem"), and compete for scarce public health resources. These issues become increasingly important as advances in interoperability and data acquisition make integrated health surveillance systems a reality. Health surveillance systems combine GIS and disease clustering software, and raise the possibility of real-time proactive disease clustering.[2] Thus the question of opportunity costs is destined to become even more pressing: is GIS a useful epidemiological tool, drawing on the technological cornucopia envisioned by Gerlernter, or is it simply a convenient way of making maps, one whose applications are ultimately limited? In particular, can the combination of GIS and disease cluster statistics increase our understanding of disease etiology? Or are health surveillance systems technologic flashes in the pan that contribute little to our understanding of human disease?

A Vision of GIS in Public Health

My vision of GIS is of an enabling technology that may lead to fundamental advances in our understanding of relationships between the environment and human health (see reference 4 for more details of this vision). The approach incorporates disease cluster statistics and other tests for spatial patterns, with the objective of generating and testing epidemiological hypotheses. This paradigm is evolving, and its potential is best understood using the water drop lens as an historical analog.

In the 1600s Anton Van Leeuwenhoek glimpsed the first images of microscopic organisms using a water drop lens. These "animalcules" were a curiosity, and no one suspected their role in infection and disease. Improvements in technology led to the compound microscope, which in the 1800s enabled Pasteur and his colleagues to reveal the link between bacteria and infection. This set the stage for major public health successes such as the eradication of smallpox. But it was the application of the technology in the context of a systematic approach that made scientific advances possible.

This analogy suggests that the promise of GIS in public health will not be realized until the technology is applied using a systematic approach such as that proposed by

[2] Reactive clustering responds to possible clusters brought forward by concerned citizens, and hence is prone to pre-selection bias. Proactive clustering surveys health event data as they are collected to identify emerging clusters.

Karl Popper. Using Popper's scientific method, a theory is inferred from observed data and falsifiable predictions are deduced from that theory. The predictions are then evaluated by experiment. When the prediction is falsified the theory is rejected. Theories may be rejected, but not proven, and predictions must be falsifiable by experiment or other means.

A related approach called "strong inference" (5) recognizes that a researcher's knowledge changes as a study progresses. Based on her/his current knowledge of the system, the researcher first formulates a set of alternative hypotheses that could explain the observed data. Systematic experiments are then designed and executed in order to exclude false hypotheses, leaving the remaining hypotheses as the only plausible explanations. During this process the researcher's knowledge base changes, and the set of alternative hypotheses may change too. Strong inference is thus a systematic approach for evaluating hypotheses in an iterative fashion.

Although they are appropriate models for laboratory studies, these systematic approaches are not directly applicable to GIS studies, since they rely on designed experiments. Spatial data typically are observational, and the processes under study often occur on a long time span that precludes experimentation. In addition, spatial systems are usually large and difficult to manipulate. This magnifies, rather than diminishes, the need for a careful and systematic approach. Despite this, we still lack a systematic approach to the application of GIS in public health. As Jacquez (4) pointed out, many health studies are prone to the "Gee Whiz" effect. This is a leap of unsupported inference that begins with the construction of thematic maps. This cartographic exercise is undertaken to visualize spatial patterns—in fact, a dramatic pattern is an important map selection criterion (why present colleagues with a map that doesn't illustrate one's point?). We are then tempted to formulate hypotheses to explain the perceived pattern. The "Gee Whiz" fallacy results: we formulate hypotheses to explain map patterns whose existence has not been demonstrated. Because maps are selected based solely on visual impact, we accept patterns without first demonstrating that they are statistically unusual; finally, hypotheses are formulated to explain patterns that may not even exist.

All of these problems can be ameliorated by making GIS part of a systematic approach that visualizes a spatial disease pattern, evaluates that pattern's statistical significance, and then generates falsifiable hypotheses that might explain the disease processes giving rise to that pattern. Building on the work of Jacquez (see Figure 2.5 in reference 4), the GeoMed project, being conducted by BioMedware and the University of Michigan, is producing a new paradigm for the analysis of spatial disease data (Figure 1). This paradigm is the result of a joint effort by Doctors Leah Estberg, Geoffrey Jacquez, Andy Long, and Mark Wilson, and is detailed in a soon-to-appear joint publication (6).

The boxes in Figure 1 labeled "Disease Data" and "Contextual Data" represent a study's data and setting. Disease data may be locations of cases and controls, disease rates, or case counts. These may or may not have been standardized, and the cases themselves may or may not have been verified, depending on the study's *context.* Contextual data define the study's setting, as do data on the environment, covariates, and confounders. The study's setting includes personnel, institutional, administrative, political, public relations, and other factors that influence how the problem is defined, how the data are collected, how the analysis is conducted, how the results are interpreted, and how interventions are selected and executed.

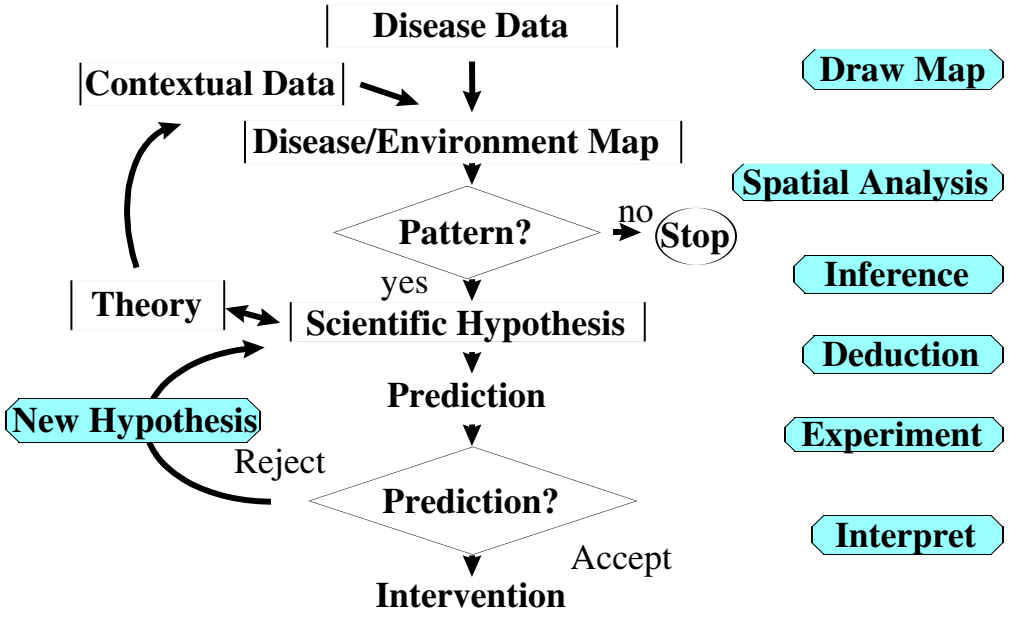

Figure 1 Spatial analysis in public health. A systematic approach for the analysis of spatial disease data.

Disease and environmental data, and information on covariates and confounders, are entered into the GIS, which is then used to construct a thematic map ("Disease/Environment Map," Figure 1) using cartographic functions such as Boolean operations, buffering, interpolation (e.g., kriging), and related techniques ("Draw Map"). The process of drawing a map is iterative and may involve several cycles through collecting data, preparing them for visualization, specifying cartographic parameters, and drawing the map. These first steps leading to map generation can be thought of as the observations in the Popperian paradigm.

Once the map is completed, the process of spatial statistical analysis begins. This process determines whether the spatial disease patterns are statistically unusual or are best explained as a chance aggregation ("Spatial Analysis"). The first step is visual inspection of the map to identify possible patterns to be statistically analyzed. Patterns of interest typically include clusters of health events and spatial associations between disease patterns and environmental variables. Both of these questions must be evaluated against the spatial fabric of disease correlates and confounders. For example, spatial clustering must be evaluated relative to the geographic distribution of the at-risk population, because population density varies from place to place. This is where disease cluster statistics and methods of spatial analysis come into play. Useful techniques include disease clustering methods, methods for analyzing spatial point distributions, adjacency statistics for determining whether classes of areas share common borders, tests for boundary overlap, statistics for evaluating association between two or more spatial variables (7–10), and related techniques for analyzing spatial data. These methods

allow us to determine whether the perceived map pattern is statistically unusual and thus warrants further investigation.

Only then can we justify inferring a theory or hypothesis to explain the spatial relationships, and proceed to the next stage ("Inference"). If the disease pattern is not significant we stop the analysis ("Stop"). This process of map generation and spatial analysis is a form of exploratory spatial data analysis, rather than a more formal approach of statistical inference. Different aspects of a spatial pattern can be explored and, hence, different statistical tests can be applied to the data. This raises issues of multiple testing, and an experiment-wise error approach, or the Bonferroni, Simes, or Holms corrections, may be needed. These techniques adjust p-values to account for repeated tests.

The decision process ("Pattern?") is based on the researcher's knowledge of the disease data and system under study (the contextual data), and *does not* proceed from the statistical results alone. It is of course possible to have significant p-values for disease patterns that are not of public health interest, as occasionally arises for cases of unrelated diseases that lack a plausible common cause or etiology. Similarly, disease patterns are occasionally found to be "not significant" even when there are compelling reasons for proceeding with the analysis.

When the map is deemed significant (e.g., when there is meaningful spatial clustering of cases above and beyond the geographic variation in density of the at-risk population), the researcher formulates a hypothesis or set of hypotheses to explain the spatial disease pattern ("Scientific Hypothesis"). The set of hypotheses may correspond to a larger body of knowledge ("Theory") describing biological mechanisms of disease causation, progression, and propagation. This body of theory contributes to the context in which the study is conducted. As hypotheses are evaluated and rejected, the underlying theory may change, giving rise to new hypotheses. This occurrence is indicated by the double-headed arrow between "Scientific Hypothesis" and "Theory."

Hypotheses in themselves are general statements that are not directly testable; a falsifiable prediction must be deduced from the hypothesis. Our next step, therefore, is to formulate a testable prediction, and design an experiment to test that prediction. As with the strong inference and Popperian approaches, we have the power only to falsify predictions and their corresponding hypotheses.

At least three kinds of experiments seem possible ("Experiment"). We may design an epidemiological study to test a prediction describing disease occurrence in populations. A laboratory study may be designed if the prediction describes disease progression at the organismic level. Finally, another GIS study may be used to evaluate epidemiological predictions that involve a spatial dimension. Notice that the GIS data used to formulate hypotheses cannot be used to test predictions that emerge from those hypotheses. To do so would bias one toward confirming the observed pattern.

Rejection of the prediction may give rise to new hypotheses, with corresponding changes in theory. Acceptance of the prediction may necessitate decisions and actions based on the experimental results. For example, a finding of a significant disease cluster, with a plausible environmental exposure as demonstrated by experiment, may warrant intervention to reduce exposure and to treat the affected population ("Intervention").

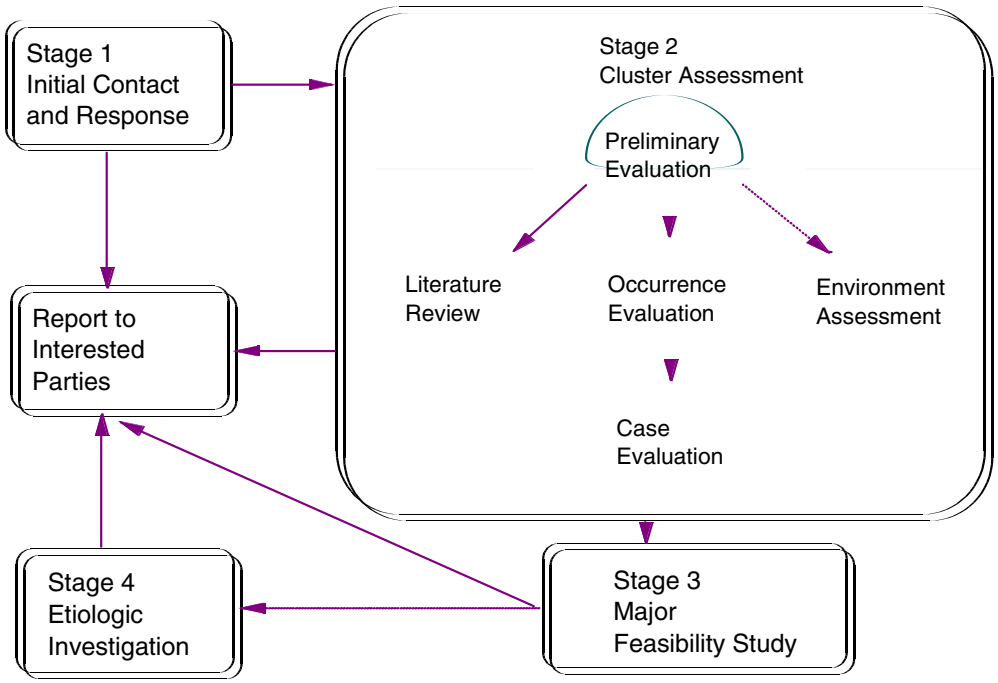

Figure 2 Disease cluster investigation protocol of the Centers for Disease Control. Modified from Centers for Disease Control, 1990 (9).

The Centers for Disease Control and Prevention's Disease Clustering Protocol

How does the schema described above relate to the Centers for Disease Control and Prevention (CDC) guidelines for investigating disease clusters? The CDC guidelines (11) advocate a four-step approach consisting of initial response, cluster assessment, major feasibility study, and finally an etiologic investigation to determine the cause of the disease cluster (Figure 2). The purpose of this protocol is to provide a systematic response to cluster allegations, to maintain good relationships with the community, and finally to conserve public health resources.

Typically, alleged clusters are brought forward by concerned citizens during stage 1, initial contact and response, during which available case data are collected. Disease cluster statistics are used in stage 2, the cluster assessment stage, primarily to determine whether a significant spatial pattern exists. If it does, more resources may be allocated for a feasibility study and etiologic investigation; if not, the investigation may be terminated. In general, disease cluster investigations are reactive and testable hypotheses are not formulated until stage 3, the major feasibility study. Only if the feasibility study is successful will an etiologic investigation take place.

Both the GIS scientific method presented earlier and the CDC guidelines use disease cluster statistics to determine whether the perceived pattern is in some sense unusual and deserving of further explanation. The CDC guidelines are thus a special case of the general protocol for spatial analysis in public health given in Figure 1.

Conclusion

Areas in which GIS technology has made substantial and continuing contributions include exposure assessment, identifying study populations, constructing disease maps and atlases, and disease surveillance to identify the locations of possible outbreaks. Other areas include the geographic placement of health services. Many of these activities yield valuable results without passing through the entire flowchart shown in Figure 1. However, whenever decisions must be made, resources must be allocated, and whenever interventions are needed, this protocol of spatial analysis in public health should be followed. In addition, a systematic approach such as that in Figure 1 must be followed if we are to advance spatial epidemiology as a scientific field by evaluating spatio-epidemiologic theories of disease spread and causation.

The approach in Figure 1 is under development in a collaborative research arrangement between BioMedware and the University of Michigan and is subject to modification. The most pressing need is an improved understanding of the relationships between space-time disease processes and the spatial disease patterns they produce. We do not expect to find a one-to-one mapping of disease process to disease pattern. However, given an observed spatial disease pattern, we do hope to be able to exclude certain disease processes as causal explanations. The 1990s experienced rapid growth in methods of spatial analysis in general and of disease cluster statistics in particular. Our arsenal of spatial analysis tools is robust and will continue to expand. In contrast, our understanding of the relationships between human diseases and their resulting spatial disease patterns is woefully inadequate. The elucidation of these relationships is the salient research need in spatial health analysis.

Acknowledgments

BioMedware and the University of Michigan, with funding from the National Cancer Institute, are preparing Web-based course materials for teaching spatial epidemiology. The graduate-level course "Spatial Epidemiology," offered at the University of Michigan School of Public Health in 1999 and 2000, has the objective of teaching the elucidation of disease processes from spatial disease patterns. The author thanks Doctors Leah Estberg, Andy Long, and Mark Wilson, who are working with the author on that project, for lively discussions on the evolving paradigm of public health surveillance. This research was funded by grants CA65366 and CA64979 from the National Cancer Institute. The views stated in this publication are those of the author, and do not necessarily reflect the perspectives of the National Cancer Institute.

References

1. Gerlernter DH. 1991. Mirror worlds: Or the day software puts the universe in a shoebox . . . how it will happen and what it will mean. Oxford University Press.
2. Marbury M. 1996. GIS: New tool or new toy? *Health and Environment Digest* 9:88–9.
3. Rothman KJ. 1990. A sobering start for the cluster buster's conference. *American Journal of Epidemiology* 132(Supplement No. 1):S6–13.
4. Jacquez, GM. 1998. GIS as an enabling technology. In: *GIS and health.* Ed. A Gattrell, M Loytonen. London: Taylor & Francis. 17–28.

5. Platt JR. 1964. Strong inference. *Science* 146:347–53.
6. Jacquez GM, Estberg L, Long A, Wilson ML. *Project GeoMed: Software and educational modules for spatial analysis in epidemiology.* Presented at the International Conference on the Analysis and Interpretation of Disease Clusters and Ecological Studies. December 16–17, 1999. Conference proceedings to appear in *Journal of the Royal Statistical Society.*
7. Jacquez GM, Waller LA, Grimson R, Wartenberg D. 1996. The analysis of disease clusters, Part I: State of the art. *Infection Control and Hospital Epidemiology* 17:319–27.
8. Jacquez GM, Grimson R, Waller LA, Wartenberg D. 1996. The analysis of disease clusters, Part II: Introduction to techniques. *Infection Control and Hospital Epidemiology* 17:385–97.
9. Haining R. 1998. Spatial statistics and the analysis of health data. In: *GIS and health.* Ed. A Gatrell, M Loytonen. London: Taylor & Francis. 29–48.
10. Kulldorff M. 1998. Selection of statistical methods for the analysis of spatial health data. *GIS and Health.* Ed. A Gatrell, M Loytonen. London: Taylor and Francis. 49–62.
11. Centers for Disease Control. 1990. Guidelines for investigating clusters of health events. *Morbidity and Mortality Weekly Report* 39:1–23.